CPSC 540 Project: A Survey on Active Learning

Setareh Cohan
Department of Computer Science
The University of British Columbia
Vancouver, BC
setarehc@cs.ubc.ca

Saeid Naderiparizi
Department of Computer Science
The University of British Columbia
Vancouver, BC
saeidnp@cs.ubc.ca

Abstract

Active learning is well-suited to many learning problems, where unlabeled data may be abundant but annotation is slow and expensive. For this work, we read a large number of papers on active learning and classified the active learning methods into four general groups. We discuss what active learning is and why it is an important area of machine learning. We then classify methods we have seen, discuss how each class of method works and, what the strengths and weaknesses of it are. We then focus on a more recent area, active learning for deep networks. More specifically, we focus on deep networks that perform image classification. We summarized a number of works we found in this area and compare them with one another. Finally, we conclude by stating what the strengths and weaknesses of active learning methods are and, what the potential future directions can be.

1 Introduction

1.1 What is active learning and why is it important?

Active learning is a subfield of machine learning where the learning algorithm is allowed to choose the data which it learns from.

For any supervised learning system to perform well, it must often be trained on large number of labeled instances. For many supervised learning tasks, labeled instances are very difficult, time consuming, or expensive to obtain. Active learning systems attempt to overcome the labeling bottleneck by asking "queries" in the form of unlabeled instances to be labeled by an "oracle" (e.g., a human expert). Thus, the active learner achieves a certain level of accuracy using fewer labeled instances, minimizing the cost of obtaining labeled data. Active learning methods have also shown to remove noisy instances from the data, which can be beneficial from an accuracy perspective. Active learning is well-motivated in many modern machine learning problems where data may be abundant but labels are scarce, expensive or, difficult to obtain.

2 Active learning literature summary

We started with reading a large number of papers on active learning and then focusing on active learning on deep neural networks for image classification task. We first give an overview of the literature by classifying active learning methods in four groups and discussing each one. Next, we move to recent papers on deep networks and carefully summarize each one.

2.1 Overview of active learning methods

The key question in active learning algorithms at any given point is, "which samples should be selected by the learning model?" Intuitively, it is advantageous to query the most "informative" or "useful" unlabeled samples. These samples can either be generated or sampled from a given distribution. There have been many proposed ways to formulate such query strategy in the literature. We have broadly

classified querying strategies in four categories: *heterogeneity-based models*, *performance-based models*, *representativeness-based models* and *hybrid models*. We will discuss each of these categories more carefully further in this section.

2.2 Heterogeneity-Based Models

The idea in these models is to learn the most heterogeneous regions. Heterogeneity is either in terms of uncertainty of labeling by the learning model, dissimilarity with the current model, or disagreement between a committee of learning models. These different techniques will be studied in this section.

2.2.1 Uncertainty Sampling

Perhaps the simplest and most commonly used query framework is uncertainty sampling introduced by Lewis and Gale (1994)[24]. In this framework, an active learner queries the instances that it is least certain how to label.
Different uncertainty measure has been used in the literature for this method. Some of these measures are listed below[31]:

- Entropy: Entropy in an information-theoretic measure that represents the amount of information needed to encode a distribution. As such, it is often thought of as a measure of uncertainty or impurity. [2] were the first to suggest using entropy as a measure of uncertainty for active learning.
- Amount of confidence: An alternative to entropy in more complex settings (e.g., multi-label classifiers instead of binary classifiers), is querying the instance whose best labeling is the least confident. One of the first works that used this measure is by [23].
- Closeness to decision boundary: Intuitively, samples that cleanly belong to a class may be informative, but less than ones that lie closer to the separation boundaries between the different classes. Therefore, in classification tasks, closeness to decision boundaries can be used to decide whether to query a sample or not. [35] were one of the first to experiment such strategy for SVMs.

Numerous uncertainty sampling techniques have been developed in the literature using these principles including [24], [23], [35], [7], [32].

2.2.2 Query-by-Committe

Another query selection approach is the query-by-committee (QBC) algorithm proposed by [33]. This approach uses a number of different learning models called the committee, which are trained on the current set of labeled instances. These learning models are then used to predict the class label of each unlabeled instance. The instance for which the learning models disagree the most is selected to be queried in this scenario.
QBC algorithms contain a committee of models along with a measure of disagreement between them. Two main approaches have been used as a measure of disagreement. The vote entropy[2] and, the KL-divergence. Some works that use this method are [33], [10], [25], [2].

2.2.3 Expected Model Change

A decision theoretic approach is selecting new instances that result in the greatest change from the current model. Specifically, instances that result in the greatest change in gradient of the objective function with respect to the model parameters is often used in the literature. The intuition behind this framework is that it prefers instances that are likely to most influence the model. This approach has been shown to work well in empirical studies, but can be computationally expensive if both the feature space and set of labelings are very large. Works on this area are [2], [32].

Heterogeneity-based models are often easy to understand and implement. However, since their goal is to identify the most unknown regions of the space (based on the current labeling), they may sometimes lead to the identification of noisy and unrepresentative regions of the data. This disadvantage has higher impact when data is more noisy.

2.3 Performance-Based Models

This class of methods attempt to directly optimize the performance of the learning model. More specifically, they look at the effect of adding the queried instances on the performance of the learning model on the remaining unlabeled instances.
There are two classes of techniques that are based on the performance which are discussed below.

2.3.1 Expected Error Reduction

These strategies estimate the expected future error that would result if some new instance is labeled and added to the labeled instances used for training. And then, it selects the instance that minimizes that expectation. This strategy was first introduced by [29] for text classification using naive Bayes. This framework requires estimating the expected future error over unlabeled samples for each query instance. It also needs to incrementally re-train the learning model for each possible labeling. This leads to a drastic increase in computational cost (expected error reduction may be the most expensive active learning framework [31]). For some model classes such as Gaussian random fields, the incremental training procedure is efficient and exact, making this approach fairly practical. For many other model classes, this is not the case. Because of this, the applications of the estimated error reduction framework have mostly only considered simple binary classification tasks. This general framework has been used in a variety of different contexts in [29], [39], [16].

2.3.2 Expected Variance Reduction

Using the result of [14], we can say that the overall generalization error can be expressed as sum of the true label noise, model bias, and variance. Of these, only the last term is highly dependent on the choice of instances selected. Therefore, it is possible to reduce the variance instead of the error.
The main advantage of this strategy over expected error reduction, is the decrease in computational requirements and the ability to express the variance in closed form, and therefore achieving greater computational efficiency.
When the learning model has d parameters and we have U unlabeled samples, the time complexity of this method is of $O(U \times d^3)$. This quickly becomes intractable for large d and U. As a solution, many methods have been proposed that approximate the variance such as [28], [17]. However, these methods are still empirically much slower than simpler query strategies like uncertainty sampling. Works on this method can be found in [6], [5].

The main advantage of performance-based over heterogeneity-based models is that they intend to improve the error behavior on the aggregated samples, rather than looking at the uncertainty behavior of the queried instances. Therefore, unrepresentative or outlier samples are avoided. However, these methods are computationally very expensive and inefficient for large models or very large dataset.

2.4 Representativeness-Based Models

These models query the data such that the acquired instances resemble the overall distribution better. This is achieved by weighting dense regions of the input space to a higher degree during the querying process. Numerous variations of this approach have been proposed, such as those in [32], [27], [25], [11].

2.5 Hybrid Models

These models combine multiple criteria for query selection to perform active learning. The queried instances need to be informative and representative. Informativeness simply means bringing new information about the feature space. Representative means the queried instance should be less likely to be outlier data. Some work that use this strategy are [9], [18], [20], [38], [8].

3 Paper Reviews

Over the past few years, deep neural networks have been very popular, and they have been used to solve many real-world problems. Applying deep neural networks to different problems follows a universal recipe; training a deep model on a very large dataset of labeled examples. However, the

need for a large number of labeled samples is rather restrictive since labeling can be expensive or difficult. One way to ease this problem is to come up with smart ways for choosing samples to label from a very large collection such that we reach a certain accuracy with minimum number of labeled samples. As we discussed before, this is the definition of active learning.
We have focused on the intersection of active learning and deep neural networks which perform image classification and, read a number of papers to this end. We have included the reviews of each of these papers later in this section.

3.1 Captcha Recognition with Active Deep Learning [34]

3.1.1 Proposed method

The problem of breaking CAPTCHA with as few samples as possible is targeted in this paper. Despite what the paper's name suggests, their method is more like a non-uniform sampling for training a network. They assume they have a set of CAPTCHA images with the ground truth text in them and they train a deep CNN on a small subset of that and iteratively add new correctly labeled samples to the training set and re-train the network. The criteria for choosing new samples to add is based on how certain the model is about its first guess in comparison with its certainty about its second guess. The network is designed to have one output neuron per each character in a CAPTCHA and they use the normalized value of network's output as the measure of model's uncertainty.

3.1.2 Evaluations

They evaluate their method on a CAPTCHA breaking task with synthetically generated samples and they compare their method with methods using different criteria for choosing new samples among correctly labeled samples, namely random, all correctly labeled, certain samples and, uncertain samples according to their proposed measure of uncertainty. They show that their method works the best among others, however, they did not compare with adding incorrectly labeled samples or using other measures of uncertainty.

3.1.3 Strengths and weaknesses

The problem of breaking CAPTCHA is interesting and fits well with the characteristics of active learning, however, their method needs a corpus of labeled training samples and the model cannot ask for labeling new samples. Also, they used the "score" network gives for each class as a measure of certainty which is not a good choice. Furthermore, they do not show why considering only correctly labeled samples makes sense rather than using incorrectly labeled ones. A major limitation of this method is that it is assumed that all CAPTCHAs have the same number of characters and characters themselves are limited to digits and English alphabet.

3.2 Cost-Effective Active Learning for Deep Image Classification [36]

3.2.1 Proposed method

Deep networks have some conflicts with active learning assumptions and settings which makes it hard to adopt active learning to deep networks. A major problem is that deep networks need a huge corpus of data while in active learning lots of data remain unlabeled and simply gets ignored. This paper is the first one to directly address using active learning for image classification via deep CNN. This paper proposes a method inspired by self-paced learning [22] and curriculum learning [3], called "Cost-Effective Active Learning" (CEAL) method to tackle these problems. The idea is that other than considering uncertain samples and asking the agent for labeling them, we can also make use of the network trained so far to get a pseudo-label for samples that network is certain about. In this way, the model will not ask for relabeling more than other methods, meanwhile, the majority high confidence samples help to learn more discriminative feature representations. Finally, for measuring uncertainty they consider least confidence (the maximum value of softmax layer on top of the network with one-hot embeddings), margin sampling (same as the criteria used in [34]) and entropy.

3.2.2 Evaluations

They evaluate their model on a face recognition task on CACD database [4] and an object categorization task on Caltech-256 [15]. They compare their method with different criteria such as random and TCAL [1] which was the state-of-the-art method but it was built to be used with SVMs. Their method outperforms all other methods and criterion and also they tried training the model having all samples labeled and used it as an upper-bound. Furthermore, they evaluate the effect of making use of high confidence samples with pseudo-labels and also different settings for choosing the confidence threshold choosing high confidence samples.

3.2.3 Strengths and weaknesses

Their idea of using high confidence samples and adding them to the training set is interesting. However, it brings another challenge of choosing the threshold. Moreover, the experiments were on relatively small datasets (around 10k-30k samples) and it might be hard to get similar results on larger datasets such as ImageNet or datasets with even larger images. Another downside of the method is that softmax probabilities are not necessarily a good measure for model uncertainty.

3.3 Deep Bayesian Active Learning with Image Data [13]

3.3.1 Proposed method

This paper uses Bayesian approach to use for adapting active learning to deep CNNs. Specifically, this paper addresses the problem of not having a measure for uncertainty in CNNs and lack of scalability in active learning methods to high-dimensional data and the problem of CNNs overfit quickly. The idea is to use Bayesian CNNs [12] which can be used to get a sensible measure of model uncertainty and also works well even with few training samples. Moreover as proven in [12] that dropout can be used to perform practical approximate inference in complex deep models, it allows to implement the model without increase in time complexity. The rest of the method is similar to other previously proposed active learning methods where a acquisition function is defined based on a criterion and new samples are chosen based on this function. In this paper, multiple acquisition functions are considered: Max Entropy, maximum information gain (as known as BALD [19]), maximum variation ratios, maximum mean STD and, random.

3.3.2 Evaluations

They do multiple experiments for evaluating different aspects of their proposed method. First of all, they compare their method with different acquisition functions and BALD, max entropy and max variation ratios almost always perform better than the others. Then, they show importance of model uncertainty (i.e. uncertainty defined in Bayesian CNNs versus taking softmax output of a CNN as uncertainty measure) by training a deterministic CNN and also a Bayesian CNN and comparing their error rate. Moreover, they compared their method with other active learning and also semi-supervised methods. Lastly, they evaluated their method on a cancer diagnosis task from lesion image data.

3.3.3 Strengths and weaknesses

This paper was the first to define a measure for real uncertainty in image classification task via deep networks which significantly improves the results. On the other hand, to get such results with Bayesian CNNs you have to reset the model after receiving each new training sample and train it again in comparison with other methods involving CNNs (e.g. [36]) which at each step only fine tunes the CNN with new samples.

3.4 Active Learning for Convolutional Neural Networks: A Core-Set Approach [30]

3.4.1 Proposed method

This paper focuses on batch active learning on convolutional neural networks. They first write an upper bound on the active learning loss such that the active learning loss is bounded by the summation of the *generalization error*, the *training error* and, the *core-set loss*. Since it is widely observed that CNNs can achieve very low training error and generalization error of CNNs is shown to be

bounded[37], they rewrite their active learning loss using only the core-set loss. This loss function is not computable since it needs all the labels. Therefore, they first give an optimizable upper bound for this objective function. To do so, they first present a bound for any Lipschitz loss function with fixed correct labels and, show that loss functions of CNNs with ReLu activations (with zero training error) satisfy this property as well. Now, the active learning problem is equivalent to k-Center problem which is NP-hard. However, it is possible to obtain a 2-optimal solution for this problem efficiently using a greedy approach. They use the same greedy approach to their problem and perform active learning to query images to labels.

3.4.2 Evaluations

They test their algorithm on the problem of image classification using three different datasets of CIFAR[21], Caltech-256[15] and SVHN[26]. They compare accuracy obtained by their method with state of the art implementations of seven active learning schemes: random sampling, uncertainty sampling, deep Bayesian active learning, best oracle uncertainty, k-Medians, batch mode discriminative-representative active learning, and CEAL. They consider both the fully-supervised and weakly-supervised settings. Fully-supervised is the case where training the classifier is done using only the labeled data points. Weakly-supervised is the case where training also utilizes the points which are not labelled yet. Their method outperforms all other methods in all experiments by a large margin.

3.4.3 Strengths and weaknesses

One of the strong points of this paper is the novelty of their approach. The paper was also very informative and detailed. Moreover, previous works all focus on fully-supervised case while this work considers both the fully and weakly supervised cases. Unlike most previous works, their method is a batch active learner. One of the major weaknesses of their method is the large number of assumptions and estimations they use (e.g., zero training error assumption) which makes their approach questionable. Another drawback is that the datasets used for experiments are rather small datasets. Also, their method is not straightforward. One potential extension of this model would be to add uncertainty to it.

4 Discussion

Active learning has been widely studied in the literature. Heterogeneity-based models are relatively easy to understand and compute, however, they might lead to choosing outliers and noisy samples. They also cannot be used with today's deep neural networks due to the correlations caused by batch sampling.
Performance-based methods are a more theoretically-motivated class of active learners. Since they directly choose samples with the goal to optimize the performance of the learner, do not suffer from adding noisy data as heterogeneity-based models do. However, they are generally computationally expensive. There have been a huge number of works suggesting optimizations for this class of methods, however, they are still more inefficient than heterogeneity-based methods and, not applicable to large datasets and complex learning models.

One trend that we observed was the focus of active learning methods on linear classifiers such as SVMs and, then shifting towards deep neural networks.

Due to the tremendous growth of data, and complexity of learning models which need annotated data for training, active learning can help reduce the labeling cost. One opportunity will be to design an active learner that is efficient for large-scale datasets and complex models. Another future direction will be defining a computationally efficient uncertainty measure such that it captures the true uncertainty of model on data. Such definition of uncertainty has already been proposed in [13] but their method is not efficient. Because of the model they use, they have to reset the model after getting new data which makes the training inefficient. Therefore, defining such a computationally efficient uncertainty measure remains an open research area.

References

[1]

[2] Shlomo Argamon-Engelson and Ido Dagan. Committee-based sample selection for probabilistic classifiers. *Journal of Artificial Intelligence Research*, 11(335):360, 1999.

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009.

[4] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *European Conference on Computer Vision*, pages 768–783. Springer, 2014.

[5] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

[6] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 1996.

[7] Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. 2005.

[8] Begüm Demir, Claudio Persello, and Lorenzo Bruzzone. Batch-mode active-learning methods for the interactive classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 49(3):1014–1031, 2011.

[9] Pinar Donmez, Jaime G Carbonell, and Paul N Bennett. Dual strategy active learning. In *European Conference on Machine Learning*, pages 116–127. Springer, 2007.

[10] Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2-3):133–168, 1997.

[11] Atsushi Fujii, Takenobu Tokunaga, Kentaro Inui, and Hozumi Tanaka. Selective sampling for example-based word sense disambiguation. *Computational Linguistics*, 24(4):573–597, 1998.

[12] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

[13] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. *arXiv preprint arXiv:1703.02910*, 2017.

[14] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

[15] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[16] Yuhong Guo and Russell Greiner. Optimistic active-learning using mutual information.

[17] Steven CH Hoi, Rong Jin, and Michael R Lyu. Large-scale text categorization by batch mode active learning. In *Proceedings of the 15th international conference on World Wide Web*, pages 633–642. ACM, 2006.

[18] Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu. Semisupervised svm batch mode active learning with applications to image retrieval. *ACM Transactions on Information Systems (TOIS)*, 27(3):16, 2009.

[19] Alex Holub, Pietro Perona, and Michael C Burl. Entropy-based active learning for object recognition. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008.

[20] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. In *Advances in neural information processing systems*, pages 892–900, 2010.

[21] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[22] M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pages 1189–1197, 2010.

[23] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine Learning Proceedings 1994*, pages 148–156. Elsevier, 1994.

[24] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.

[25] Andrew Kachites McCallumzy and Kamal Nigamy. Employing em and pool-based active learning for text classification. Citeseer.

[26] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 5, 2011.

[27] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79. ACM, 2004.

[28] Gerhard Paass and Jörg Kindermann. Bayesian query construction for neural network models. In *Advances in Neural Information Processing Systems*, pages 443–450, 1995.

[29] Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, pages 441–448, 2001.

[30] Ozan Sener and Silvio Savarese. A geometric approach to active learning for convolutional neural networks. *arXiv preprint arXiv:1708.00489*, 2017.

[31] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[32] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1070–1079. Association for Computational Linguistics, 2008.

[33] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM, 1992.

[34] Fabian Stark, Caner Hazırbas, Rudolph Triebel, and Daniel Cremers. Captcha recognition with active deep learning. In *Workshop New Challenges in Neural Computation 2015*, page 94. Citeseer, 2015.

[35] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[36] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2017.

[37] Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

[38] Zhao Xu, Kai Yu, Volker Tresp, Xiaowei Xu, and Jizhi Wang. Representative sampling for text classification using support vector machines. In *European Conference on Information Retrieval*, pages 393–407. Springer, 2003.

[39] Xiaojin Zhu, John Lafferty, and Zoubin Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions.